Advantage of Persistent Memory from Operational Perspective

Northern California Oracle Users Group – Summer 2021

Speakers

Vincent Chong

- Sr. MTS Engineer
- Database Cloud Engineering

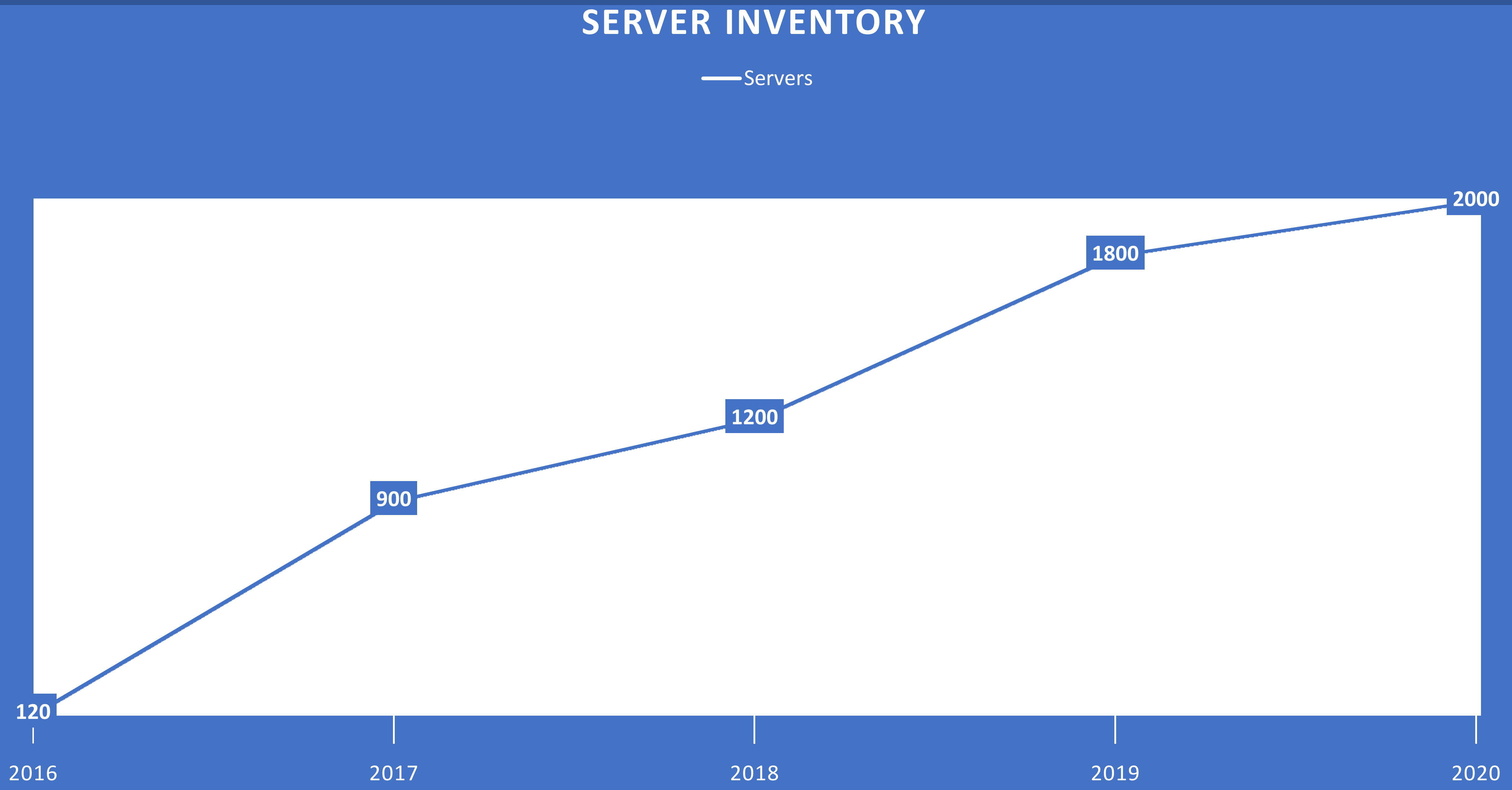
Deepti Sharma

- MTS Engineer
- Database Cloud Engineering

Content

- Aerospike Footprint at PayPal
- Aerospike DB Architecture
- Persistent Memory Advantage
- Aerospike HW Configuration at PayPal
- Operational Benefits of PMEM

Aerospike - 5 Years Journey

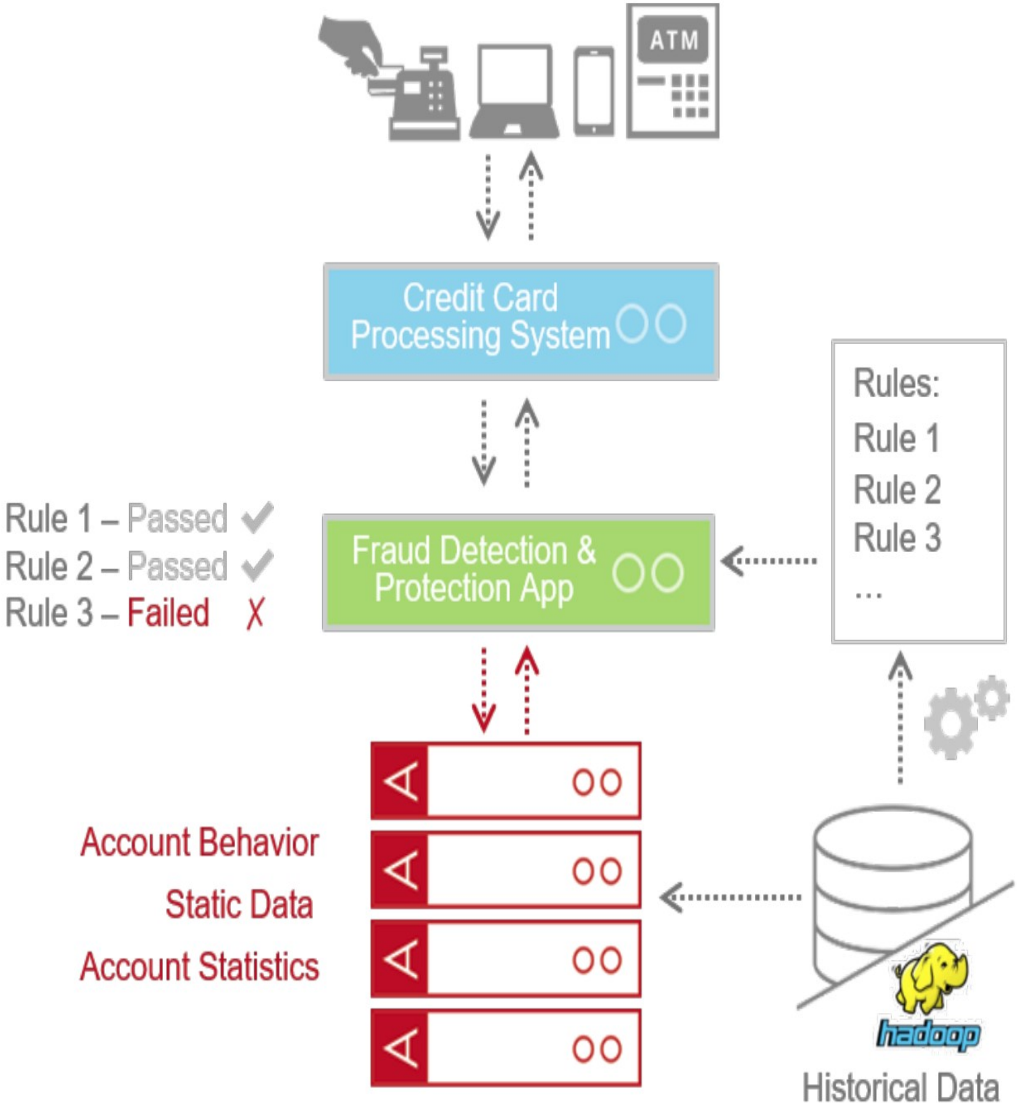


Use cases

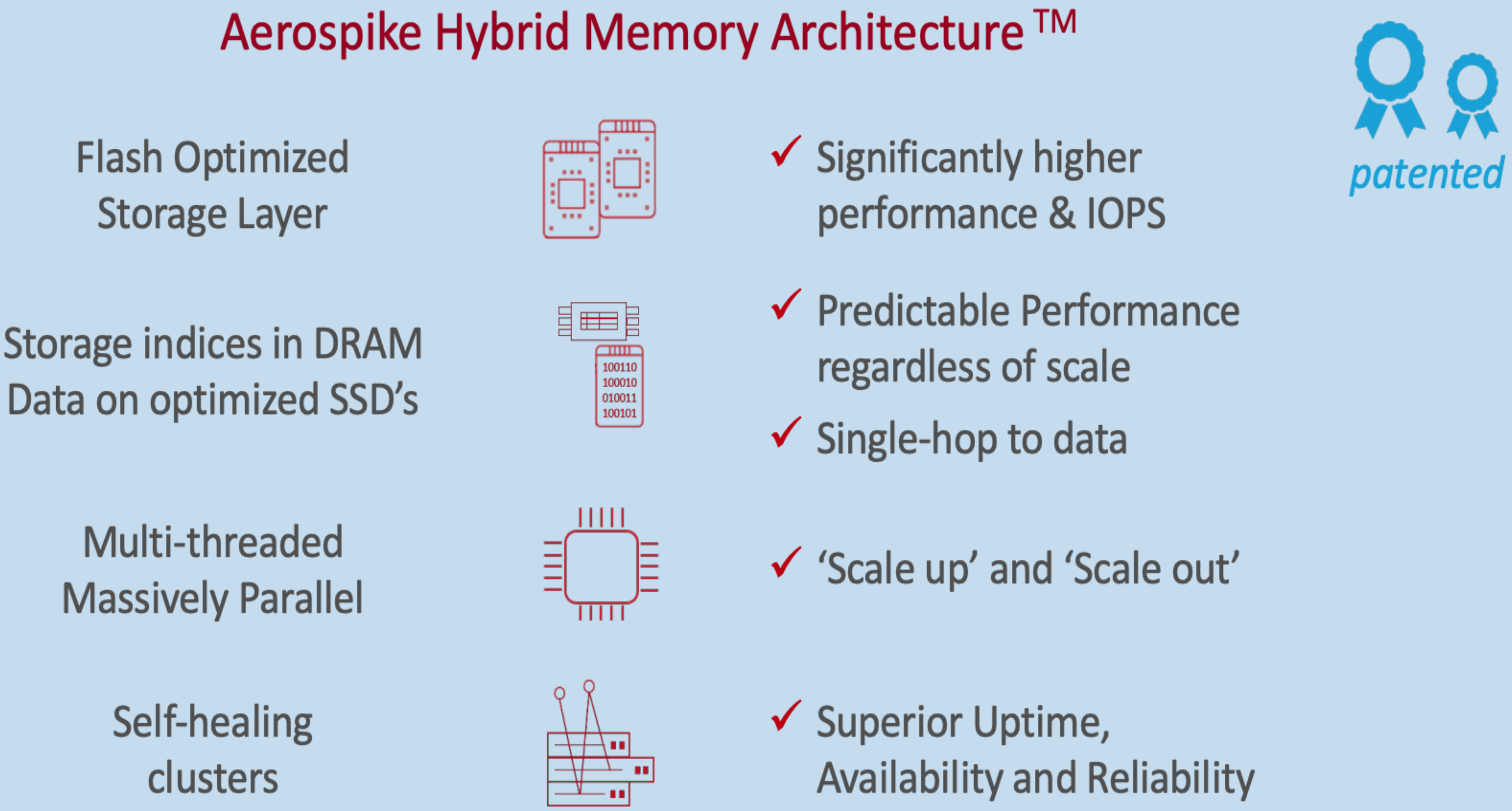
- Fraud detections
- Compliance checks
- Graph relationships
- Mobile device fingerprints
- Event histories

Fraud Detection

- Powering Global Fraud Prevention Network
  - $280 B Payments annually
- Replaced Terracotta Server Array Cache
  - Reduced erver footprint by 15x
- Improved SLAs
  - 30x reduction in false positives
- Increased revenue
  - 10x improvement in fraud calculation data used



Hybrid Memory Architecture



Aerospike Hybrid Memory Architecture ^{TM}

patented

Flash Optimized
Storage Layer

✓ Significantly higher performance & IOPS

Storage indices in DRAM
Data on optimized SSD's

100110
100010
010011
100101

✓ Predictable Performance regardless of scale

✓ Single-hop to data

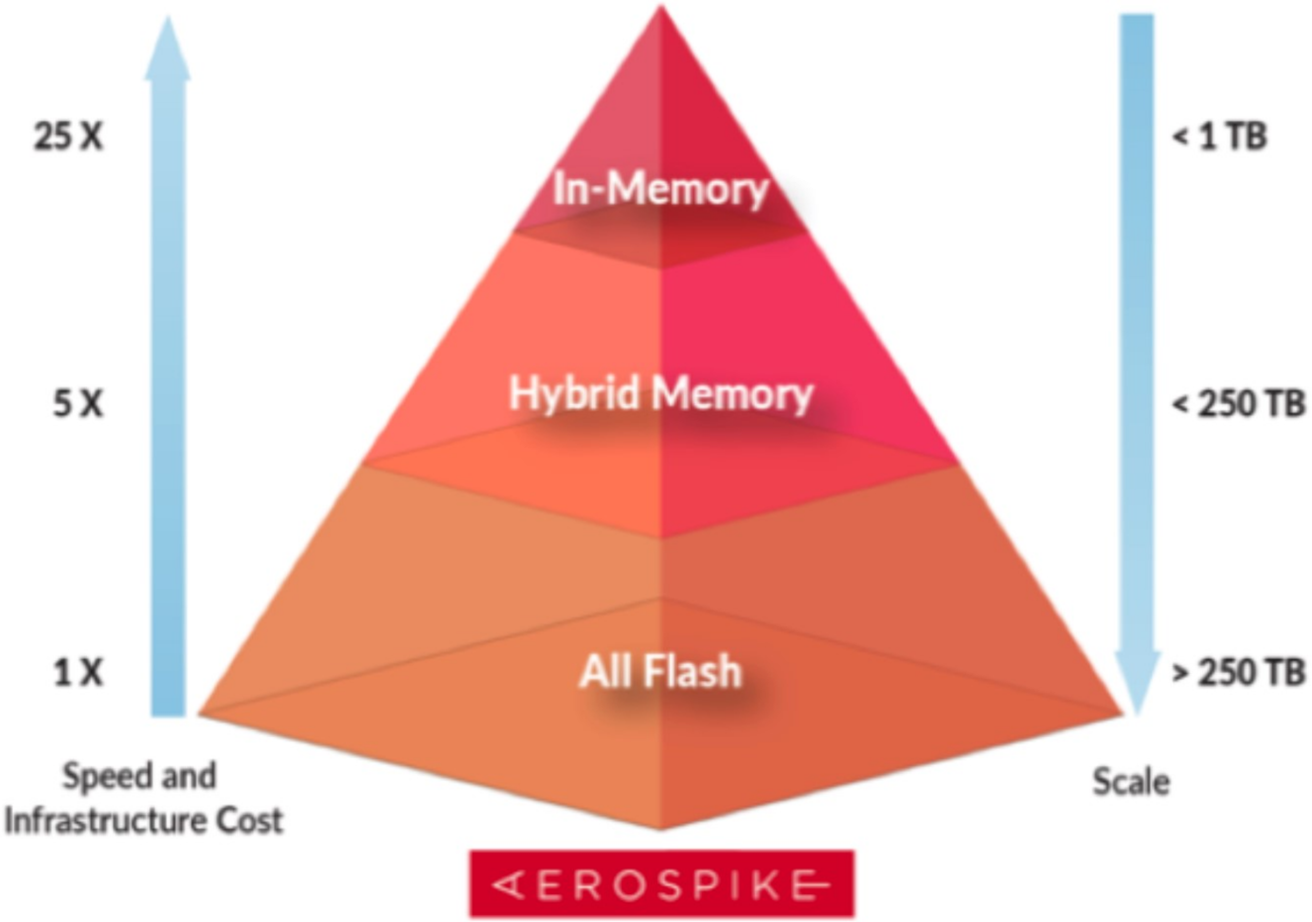
Multi-threaded
Massively Parallel

✓ 'Scale up' and 'Scale out'

Self-healing
clusters

✓ Superior Uptime, Availability and Reliability

Hybrid Memory Architecture



Aerospike's Tiered Architecture is unique

- In-Memory
- Hybrid Memory
- All Flash

Main Challenges

- Growing cost of supporting data volume growth
- Maintaining performance
- Operational efficiency
    - Time to failure detection
    - Time to recovery
    - Compliance with security requirements
      e.g. OS patching

Solution – Persistent Memory

- A type of non-volatile media that fits in a standard DIMM (memory) slot
- It's slower than DRAM, but provides higher throughput than NVMe SSD
- Much larger capacities than DRAM and are less expensive per GB
- Still more expensive than NVMe SSD

Identical form factor as DRAM

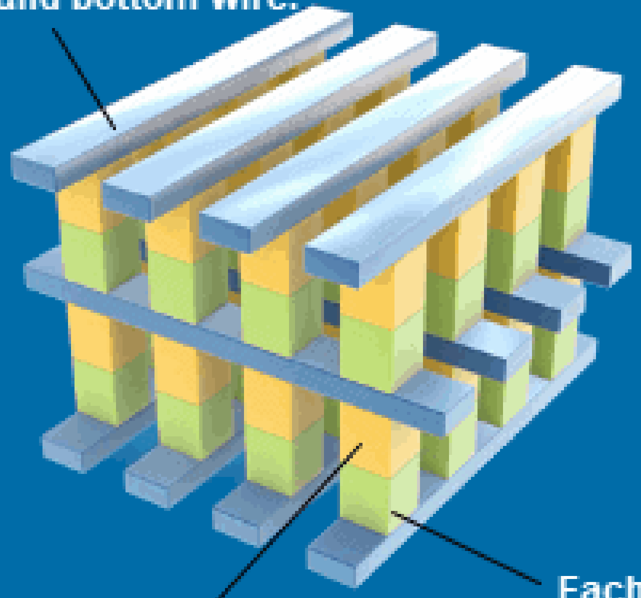
Access methods

Block access, which operates like storage for app compatibility. In this configuration, data flows through the file system and storage stacks as normal.

Direct access (DAX), which operates like memory to get the lowest latency. You can only use DAX in combination with NTFS.

Solution – Persistent Memory

3D XPoint Structure



Perpendicular wires connect columns. An individual memory cell is addressed by selecting it's top and bottom wire.

A Selector enables it's memory cell to be written/read to without a transistor.

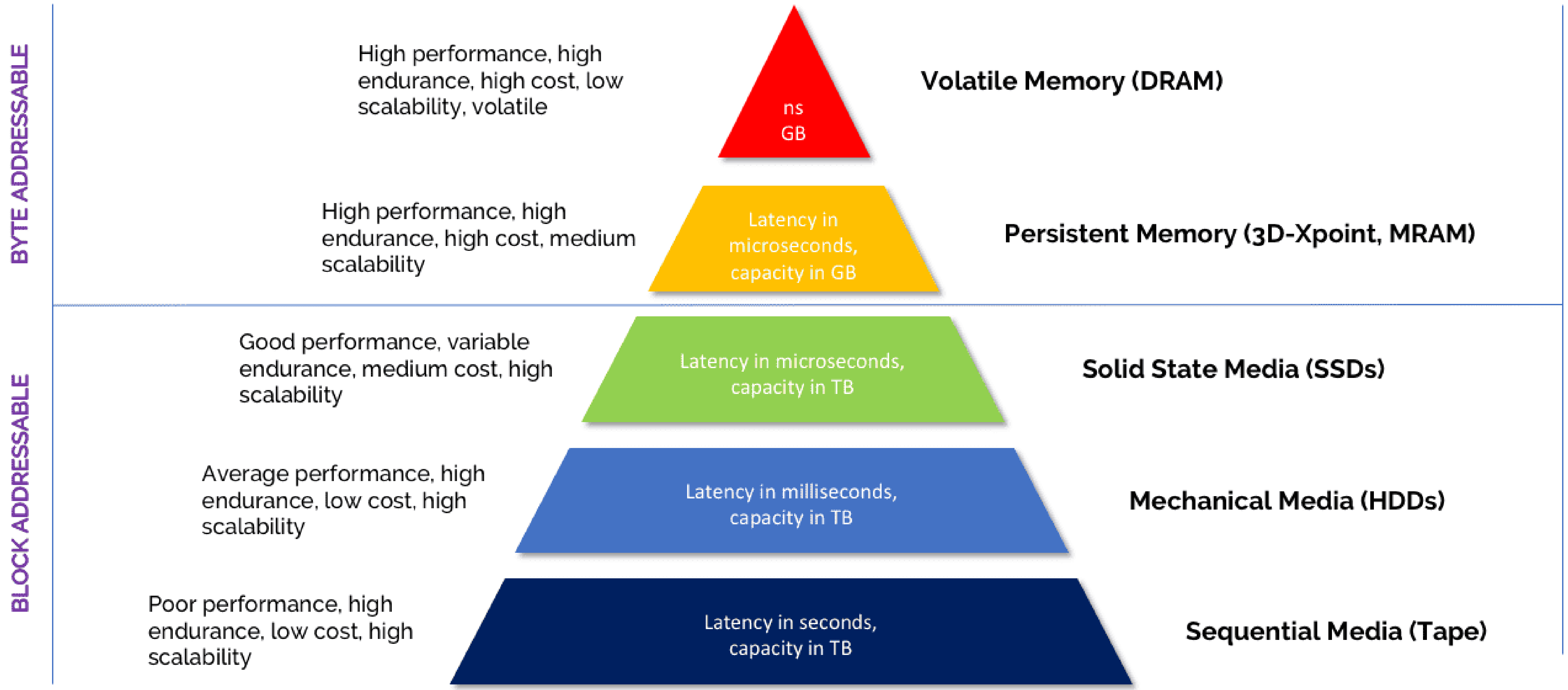
Each memory cell stores a single bit of data

Intel Optane Memory

How does PMEM works?

- Different storage physics: threshold switch, not transistor
- A bit is accessed by a current sent through the top and bottom wires touching each cell
- Cells can be stacked in three dimensions for higher capacity
- The cell can occupy either a high- or low-resistance state, representing a 1 or a 0
- Resistance state hold their values indefinitely, even when there is a power loss.
- For write operations, a specific voltage changes the resistance property of the selected cell
- For read operations, a different voltage is sent through to determine whether the cell is in a high- or low-resistance state

THE STORAGE MEDIA HIERARCHY



BYTE ADDRESSABLE

High performance, high endurance, high cost, low scalability, volatile

ns
GB

Volatile Memory (DRAM)

High performance, high endurance, high cost, medium scalability

Latency in microseconds, capacity in GB

Persistent Memory (3D-Xpoint, MRAM)

BLOCK ADDRESSABLE

Good performance, variable endurance, medium cost, high scalability

Latency in microseconds, capacity in TB

Solid State Media (SSDs)

Average performance, high endurance, low cost, high scalability

Latency in milliseconds, capacity in TB

Mechanical Media (HDDs)

Poor performance, high endurance, low cost, high scalability

Latency in seconds, capacity in TB

Sequential Media (Tape)

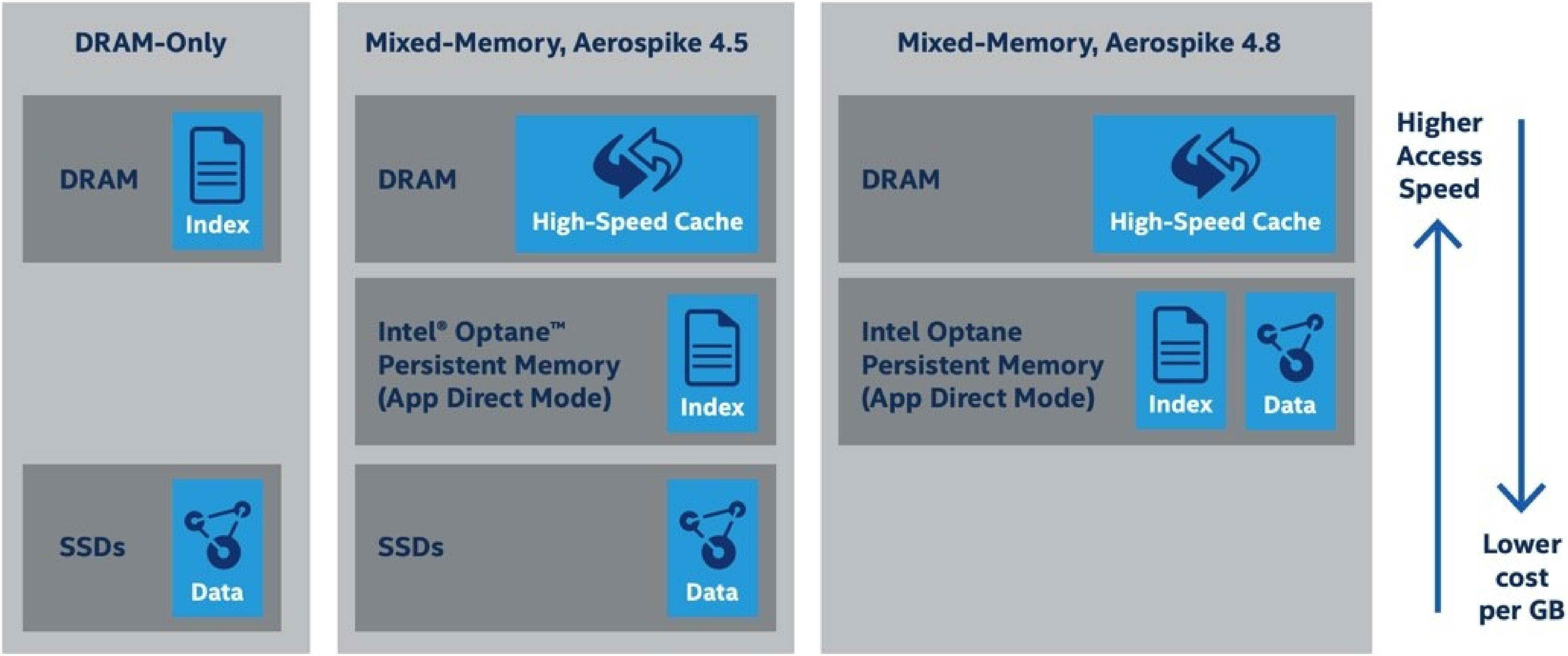
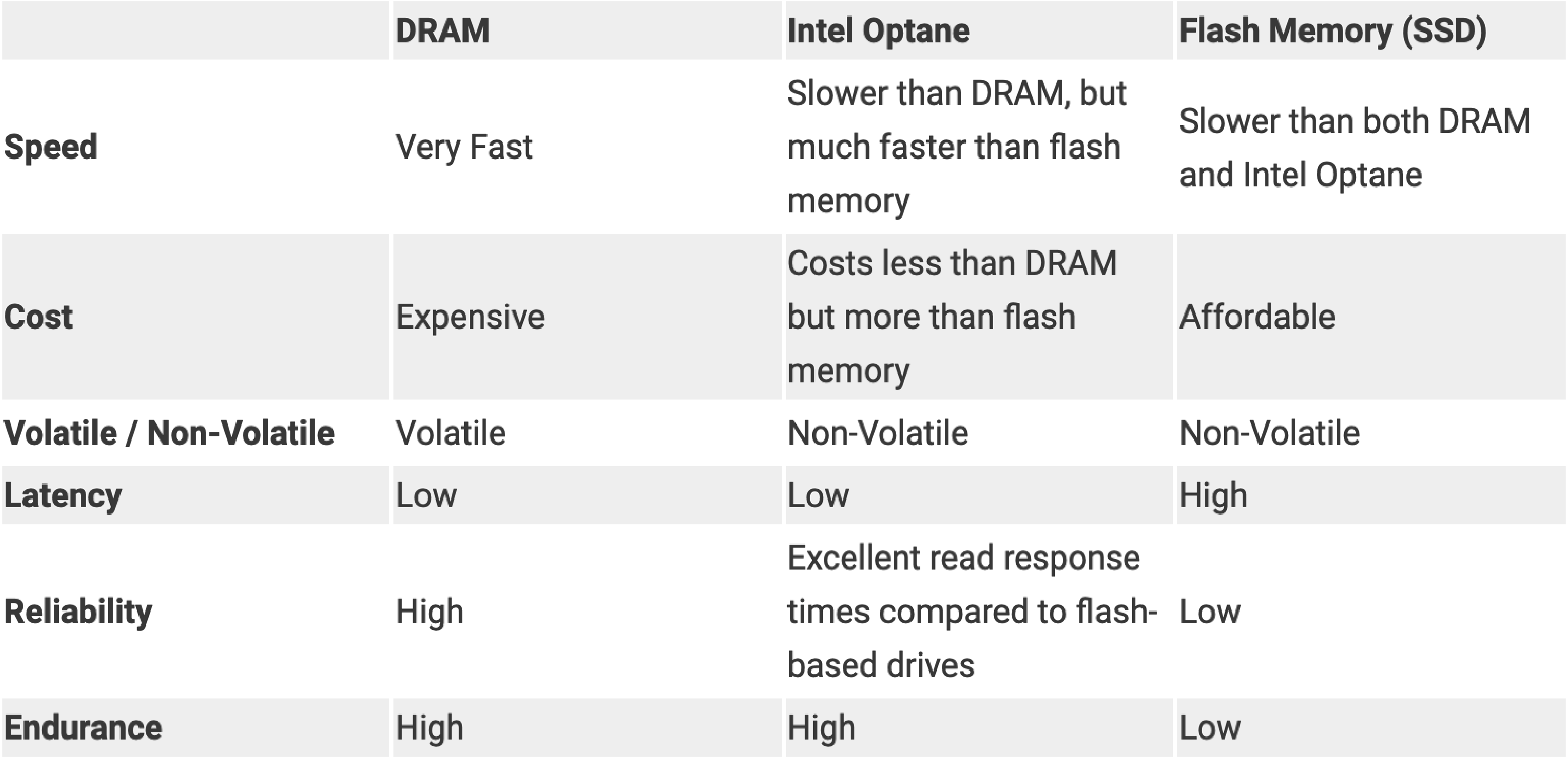
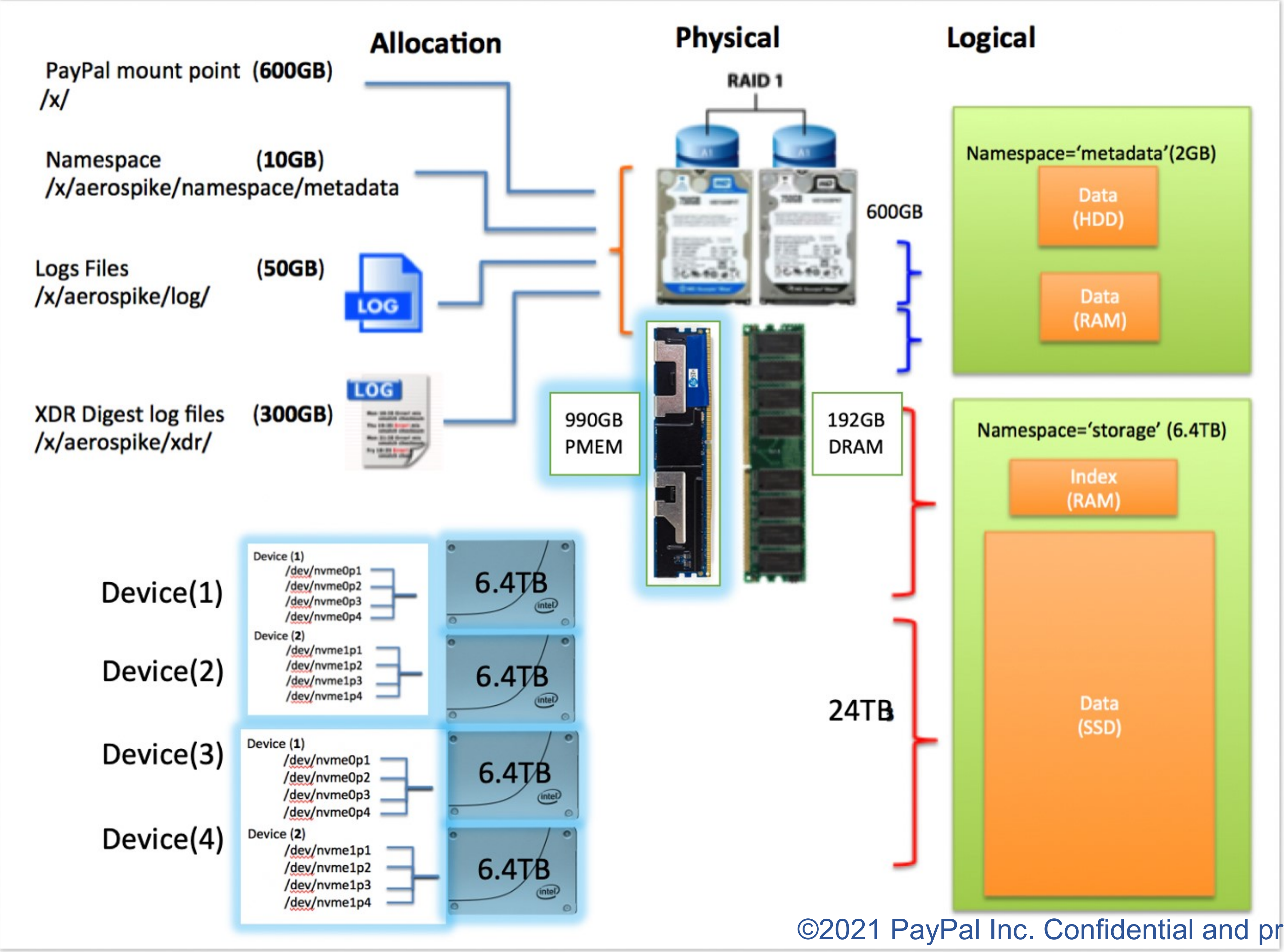

DRAM-Only
DRAM
Index
SSDs
Data
Mixed-Memory, Aerospike 4.5
DRAM
High-Speed Cache
Intel® Optane™ Persistent Memory (App Direct Mode)
Index
SSDs
Data
Mixed-Memory, Aerospike 4.8
DRAM
High-Speed Cache
Intel Optane Persistent Memory (App Direct Mode)
Index
Data
Higher Access Speed
Lower cost per GB

Figure 4. Evolution of Aerospike and Intel collaboration.

PMEM Vs SSD Vs DRAM: Performance Comparison



	DRAM	Intel Optane	Flash Memory (SSD)
Speed	Very Fast	Slower than DRAM, but much faster than flash memory	Slower than both DRAM and Intel Optane
Cost	Expensive	Costs less than DRAM but more than flash memory	Affordable
Volatile / Non-Volatile	Volatile	Non-Volatile	Non-Volatile
Latency	Low	Low	High
Reliability	High	Excellent read response times compared to flash-based drives	Low
Endurance	High	High	Low



Allocation

PayPal mount point (600GB)
/x/

Namespace (10GB)
/x/aerospike/namespace/metadata

Logs Files (50GB)
/x/aerospike/log/

XDR Digest log files (300GB)
/x/aerospike/xdr/

Physical

RAID 1

600GB

990GB
PMEM

192GB
DRAM

24TB

Logical

Namespace='metadata'(2GB)

Data
(HDD)

Data
(RAM)

Namespace='storage' (6.4TB)

Index
(RAM)

Data
(SSD)

Device(1)

Device(2)

Device(3)

Device(4)

Device (1)
/dev/nvme0p1
/dev/nvme0p2
/dev/nvme0p3
/dev/nvme0p4

Device (2)
/dev/nvme1p1
/dev/nvme1p2
/dev/nvme1p3
/dev/nvme1p4

Device (1)
/dev/nvme0p1
/dev/nvme0p2
/dev/nvme0p3
/dev/nvme0p4

Device (2)
/dev/nvme1p1
/dev/nvme1p2
/dev/nvme1p3
/dev/nvme1p4

Aerospike Footprint - 2021

- 1300 Servers (down from 2600 servers old SKU)
- ~ 44 Racks (down from 60 racks)
- > 150 Clusters
- 10s teams
- > 10K clients app
- Supports 500+ developers

Aerospike Database Indexing with pmem [in seconds]

- Storage Architectures (Recap)
- Index during maintenance operations
- Re-indexing with DRAM vs. Persistent Memory
- Stats with Persistent Memory
- Overall Gains using Persistent memory

Storage Architectures

Storage Architectures:

- In-Memory
- All Flash
- Hybrid (Memory/Flash)

Challenge:

- Reboot times

- **Index Content**

INDEX = Digest (hash) + Write generation + Expiration time + Last update time + Storage address

*Digest – Fixed 20 bytes distributed hash representing a key.

- **Storage Space needed for Index**

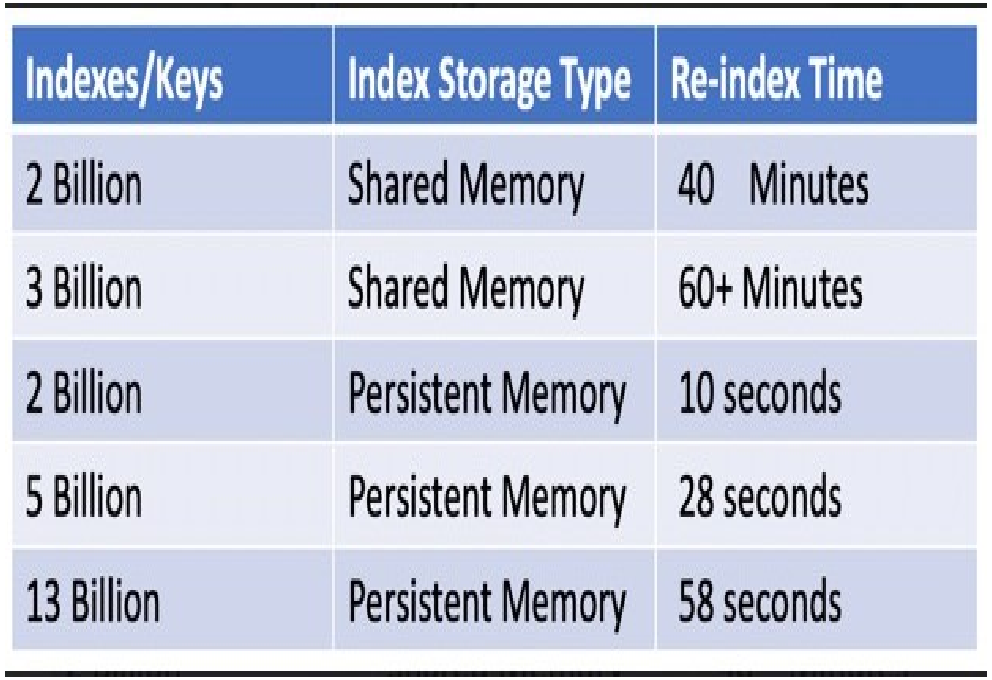
INDEX = 64 BYTE data structure = 512 total addressable BIT(s) in DRAM per key.

- **Index Persistence**

Primary index is derived from the data itself and can be rebuilt from that data.

PayPal

Index during maintenance operations



Indexes/Keys	Index Storage Type	Re-index Time
2 Billion	Shared Memory	40 Minutes
3 Billion	Shared Memory	60+ Minutes
2 Billion	Persistent Memory	10 seconds
5 Billion	Persistent Memory	28 seconds
13 Billion	Persistent Memory	58 seconds

Aerospike EE stores indexes in Linux shared memory (shmem)

But during OS reboots, Aerospike loses the indexes from shared memory (DRAM) and has to rebuild from Disk

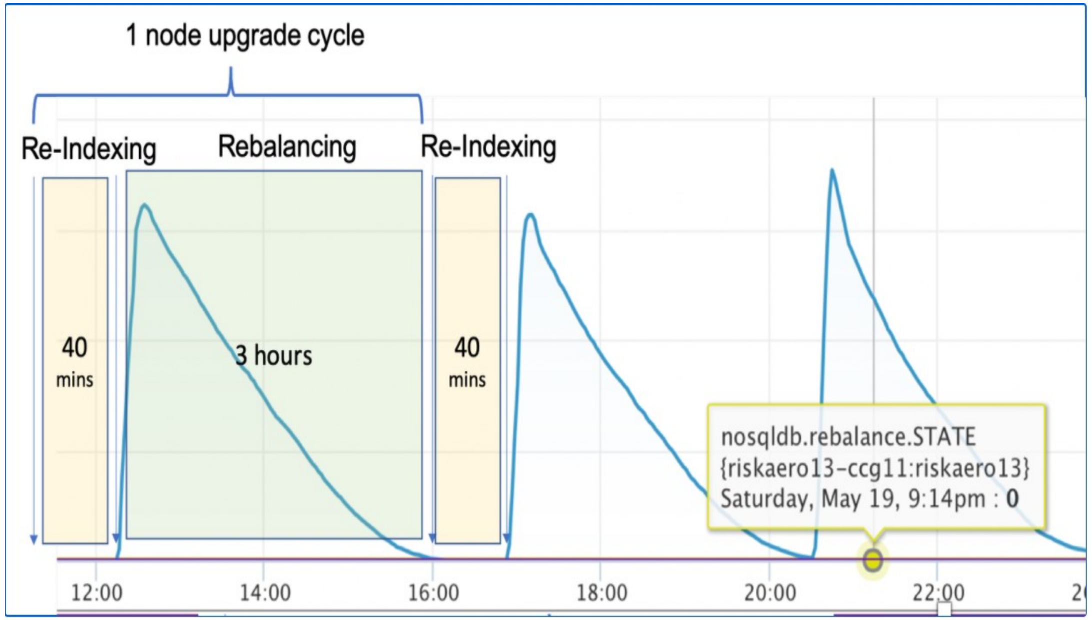
The re-indexing time for typical 2B keys from Disk is ~**40 minutes**

In a worst case, if there are needs to do full rebalance, then the time taken for full rebalance is on average 3-4 hours

With **Persistent Memory, full Indexing** during reboots **~10 seconds**.

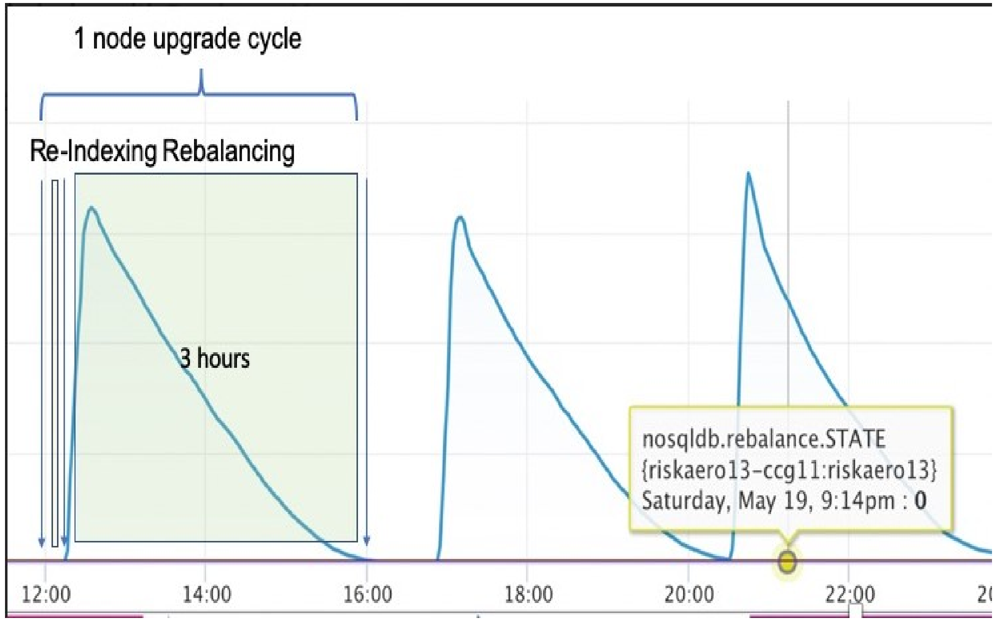
Reindexing with DRAM vs Persistent memory

Re-indexing + Full rebalance: ~3-4 hours avg depending on data density.



nosqldb.rebalance.STATE
{riskaero13-ccg11:riskaero13}
Saturday, May 19, 9:14pm : 0

Full Indexing during reboots ~10 seconds.



nosqldb.rebalance.STATE
{riskaero13-ccg11:riskaero13}
Saturday, May 19, 9:14pm : 0

OS patching improvements with persistent memory over DRAM

Total Cluster Groups: 56 (3 clusters each group)

Availability Zones: 3 - (Primary/LDR/DR)

Number of Servers: 1700

Number of Servers	DRAM (OS patching Duration)	PMEM (OS patching Duration)
One Server	1-2 hours	30-45 min
One Cluster (10 nodes)	10-12 hours	5-6 hours
1700 Servers	1700 hours (~75 days)	850 hours (~36 days)

TIME SAVED for the ENTIRE AEROSPIKE INVENTORY 2x

Environment Stats

	AerospikeSD	AerospikeHD	Improvement
Max Keys/Node	~2 Billion	~10 Billion	5x
Max Usable Storage/Node	3.2 TB	12.8 TB	4x
Nodes/Cluster	20	10	2x
Rack/Power/Space	20U	10U	2x
~Cost/ClusterGroup	~1.3M	~900K	30+% drop in price
Replication factor	2	3	Yes
Clusters/Rack	2	4	2x
ReIndexing Time	59 minutes	4 minutes	12x
Reboot+Reindex Time	1 hour	8 minutes	8x
Rolling software upgrade	10+ hours	~5 hours	Yes
Rolling OS upgrade	10+ hours	~6 hours	Yes
Cluster Creation	11 minutes	11 minutes	No

Gains with Persistent Memory

1 Opportunities

- **Data Center Migration (DCX) :** New cluster build outs with pmem configuration.
- **OS Patching:** OS vulnerabilities are becoming #1 priority for any financial domain organization which requires patching the servers as soon as new kernel version is available.

PayPal Infrastructure Patching Cycles

2 The Solution

Persistent memory configuration in Production:

- Pmem provided two-fold benefit: memory + storage
- Faster Re-indexing.
- High density storage

Speed of Execution

3 Outcomes

Faster Patching cycles within SLA (30 days)

Storage optimization

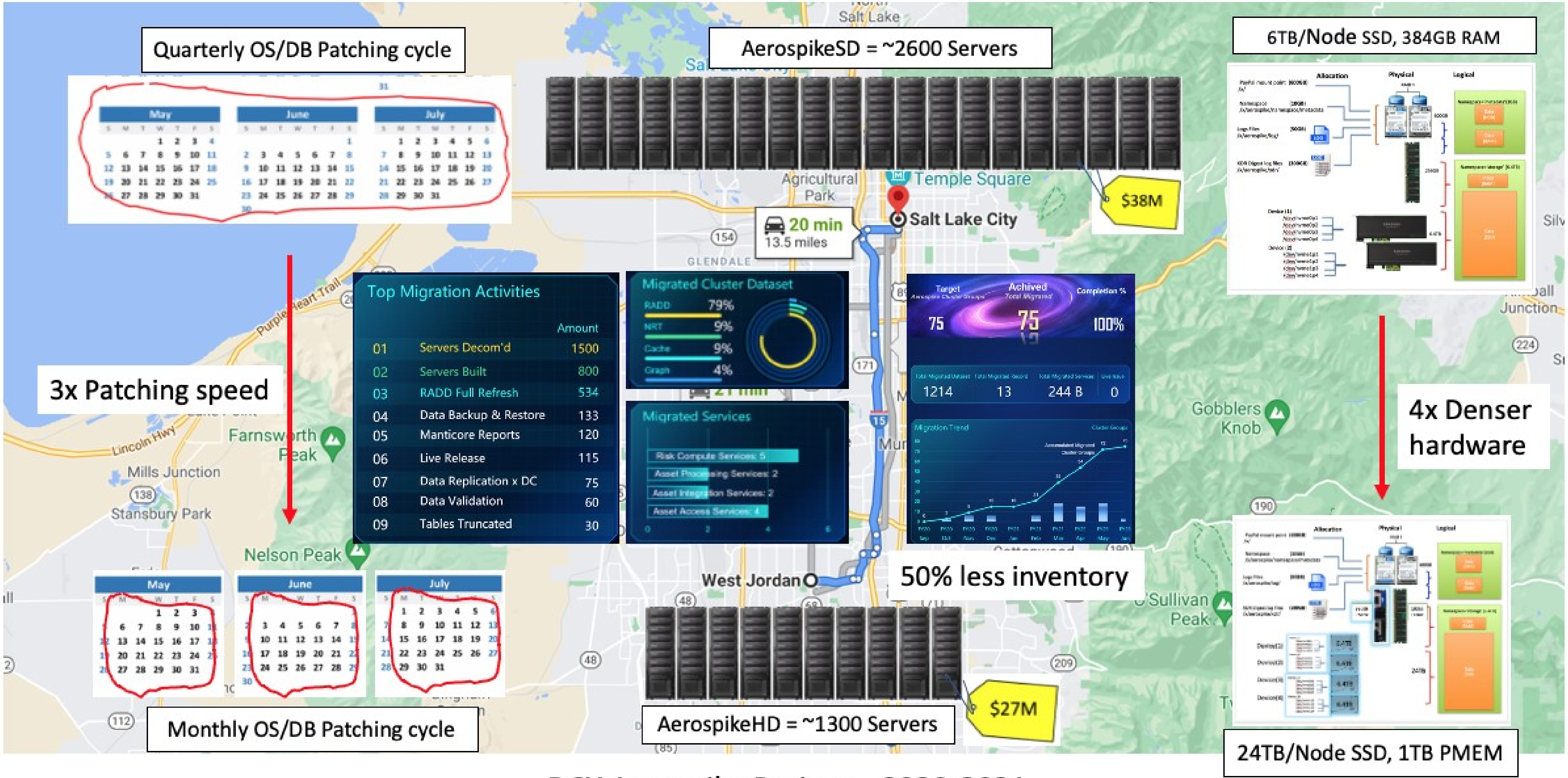
Better Performance

Lesser Cost

Operational Efficiency

PayPal

75 clusters migrated, 244B keys replicated, 1214 tables migrated, 800 servers built, 1500 servers decommissioned - **0 ATB Impact**



Quarterly OS/DB Patching cycle

AerospikeSD = ~2600 Servers

$38M

6TB/Node SSD, 384GB RAM

3x Patching speed

Top Migration Activities

		Amount
01	Servers Decom'd	1500
02	Servers Built	800
03	RADD Full Refresh	534
04	Data Backup & Restore	133
05	Manticore Reports	120
06	Live Release	115
07	Data Replication x DC	75
08	Data Validation	60
09	Tables Truncated	30

Total Migrated Dataset	Total Migrated Record	Total Migrated Services	Live Issue
1214	13	244 B	0

4x Denser hardware

50% less inventory

Monthly OS/DB Patching cycle

AerospikeHD = ~1300 Servers

$27M

24TB/Node SSD, 1TB PMEM

DCX Aerospike Project – 2020-2021

Conclusion

PERFORMANCE AND COST ARE NOT THE ONLY IMPORTANT CONSIDERATION OF USING PMEM

OPERATIONAL CONSIDERATION HAS FAR-REACHING CONSEQUENCE IN DATA RECOVERY TIME AND OPERATIONAL EFFICIENCY

Thank You !

Questions??